Multimodal Modeling of Task-Mediated Confusion

Camille Mince
Pomona College
cima2018@mymail.pomona.edu

Skye Rhomberg
Colby College
sorhom22@colby.edu

Cecilia O. Alm **Reynold Bailey** **Alexander Ororbia**
Rochester Institute of Technology
{coagla, rjbvcs, agovcs}@rit.edu

Abstract

In order to build more human-like cognitive agents, systems capable of detecting various human emotions must be designed to respond appropriately. Confusion, the combination of an emotional and cognitive state, is under-explored. In this paper, we build upon prior work to develop models that detect confusion from three modalities: video (facial features), audio (prosodic features), and text (transcribed speech features). Our research improves the data collection process by allowing for continuous (as opposed to discrete) annotation of confusion levels. We also craft models based on recurrent neural networks (RNNs) given their ability to predict sequential data. In our experiments, we find that text and video modalities are the most important in predicting confusion while the explored audio features are relatively unimportant predictors of confusion in our data.

1 Introduction

Humans are adept at recognizing the emotions of others. They can identify whether another person has positive, negative, neutral, or more nuanced emotions by considering their facial expressions, voice, and words. To construct more human-like cognitive systems, it is important that, just as humans do, computational systems can infer emotions of the users that they interact with. Modeling confusion is relatively under-explored and can be difficult to detect computationally. Confusion can occur when someone does not know how to proceed with a task or when reconciling old beliefs with confounding information. The American Psychological Association's Dictionary of Psychology defines confusion as "a mental disturbance characterized by bewilderment, inability to think clearly or act decisively, and disorientation for time, place, and person" (Association, 2021). Potential applications of a confusion-detecting agent include task-driven dialogue chat-bots and detecting a learner's confusion in online learning environments.

We present models that leverage data across several modalities - facial expressions, speech signals with prosody, and transcribed spoken language - that not only can be used to predictively model confusion but also to extract insights with respect to which features of which modalities are clearer indicators of confusion. In this work, we answer the following research questions:

RQ1 How can we improve upon prior data collection methods to obtain a more precise multimodal dataset with confusion labels?

RQ2 How accurate of a model can we construct that classifies the degree of confusion at different points within a task?

RQ3 What facial, audio, and language features serve as good predictors of confusion (or a lack thereof)?

2 Related Work

Detecting confusion has mostly been explored in educational settings to discern students' confusion. As MOOCs (Massively Open Online Courses) have become more prevalent, researchers have focused on building models that accurately detect students' confusion. Defining a learner's confusion as "an individual state of bewilderment and uncertainty as to how to move forward," Atapattu et al. (2020) found that linguistic-only features were highly accurate predictors of confusion. Using a dataset of nearly $30,000$ anonymous posts from Stanford's MOOC discussion forum, they used natural language processing resources, e.g., sentiment analysis, and a MANOVA test to extract feature importance. While Atapattu et al. (2020) focused on linguistic-only features, Shi et al. (2019) analyzed facial expressions to classify learners' confusion. They used statistical learning models that leveraged a combination of histogram of oriented gradients (HOG) features and local binary patterns (LBPs)

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 188 - 194

in tandem with a prediction system, composed of a support-vector machine (SVM) and a convolutional neural network (CNN). The CNN-SVM had the best performance, indicating that facial expressions can be good predictors of confusion.

Our research differs from these past experiments in that we aim to create a multimodal model. Furthermore, we incorporate additional speech analysis to craft a more richly informed predictor of confusion. The study most closely related with our own is Kaushik et al. (2021), which experimented with a random forest classification scheme applied over discrete time intervals extracted from two-person interactions. Notably, this work considered interpretable metrics such as disfluencies (like *um*), questions, and pauses, although these were less correlated with confusion than the best-correlated facial expressions. We expand upon the study by Kaushik et al. (2021), repeating the human subject set-up of two people collaboratively solving a task over Zoom. While that study had participants label their level of confusion across a 30-second interval, our research explores continuous annotation instead of discrete spans. We expect that continuous confusion labels will enable more useful reference data for classification.

3 Methodology

3.1 Data Collection

In this IRB-approved study, subjects were recruited through email to participate in a "conversational behavior study." We did not debrief participants until after the study was complete that the true aim was to analyze confusion. Participants were paired by availability to work together through a series of three confusion-evoking tasks. We had participants complete the tasks in pairs to elicit intuitive and meaningful interactions. Our goal was to construct a dataset of multimodal text, speech audio, and video-based facial expression features with confusion-inducing tasks. Additionally, we sought to improve upon the prior research of Kaushik et al. (2021) by supporting continuous annotation of confusion levels by participants. The first and third tasks were adapted from Kaushik et al. (2021); in the first task, participants were given four minutes to find a 30 minute meeting time given two calendars which actually had no overlapping availability (see Figure 1).

The second and third tasks were logic puzzles (one was the widely known puzzle titled "Cheryl's Birthday"). Given a list of potential birthdays and clues about which of those dates could not have been Cheryl's birthday, participants were asked to reason through hints, rule out dates, and determine Cheryl's true birthday.[1] Participants were given four to seven minutes to solve the riddles with periodic hints sent via Zoom chat. After participants completed the three tasks, they were then told that the true purpose of the experiment and asked to annotate their confusion levels throughout each task utilizing our website: the *Confuse-o-Meter*. The website displayed the playback of the participants solving the task on top of a set of radio buttons with the following labels: *Not Confused, Slightly Confused, Very Confused,* and *Extremely Confused*.

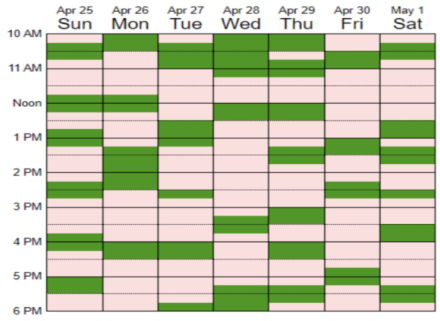

Figure 1: In one task, participants were given two calendars without overlapping availability. They were asked to find a 30 minute meeting time at which they were both available.

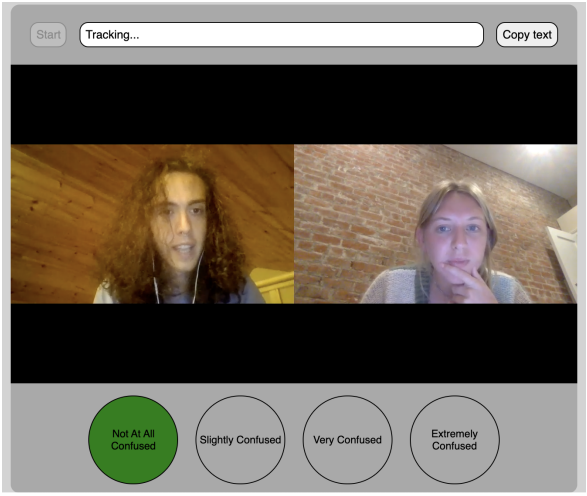

Figure 2: In our continuous confusion annotation, participants were instructed to use the four radio buttons to continuously annotate their confusion levels. They were instructed to change the radio button whenever they noticed a change in their own confusion level.

[1] https://en.wikipedia.org/wiki/Cheryl%27s_Birthday

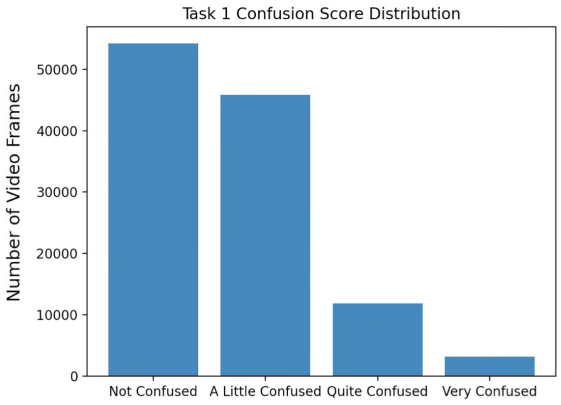

Figure 3: Our method for continuous annotation allowed for us to label each video frame with the participant's indicated confusion level. Above is the distribution of labels for each video frame. Task 1 was the least confusing task, as shown by the fact that the majority of the labels appear in the *Not Confused* state.

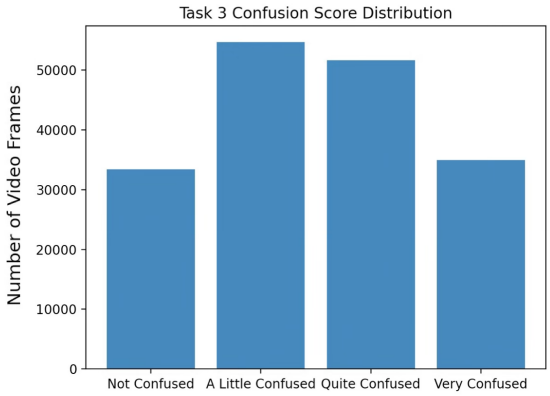

Figure 4: Task 3 was the most confusing task. Although the majority of the labels fall under *A Little Confused,* there are a considerable number of video frames labeled as *Quite Confused* and *Very Confused.*

As seen in Figure 2, participants were instructed to click on the appropriate radio button whenever they noticed a change in their confusion level. From the website, we obtained data in the following form: [(timestamp of change, new confusion label), (timestamp of next change, new confusion label),. . .]. We used this encoding to produce confusion labels for every time-step of data. This approach allowed us to generate a dataset in which every time-step of data was accurately labeled with the participant's confusion level.

Some tasks were harder than others. We intended for the tasks to progress in difficulty so that we could collect ample *Not Confused* and *Confused* data. The distribution of the participants' confusion ratings in the first and last task are shown in Figures 3 and 4, respectively. It is clear that the participants found the first task to be less confusing, with the majority of the labels being in the *Not Confused* state. A higher proportion of the labels fall in the *Quite Confused* and *Very Confused* categories for tasks 2 and 3.

3.2 Feature Extraction

The *OpenFace* (Baltrušaitis et al., 2015, 2018) software package was used to extract 17 different Ekman and Friesen (1976) facial action units (FACs) defined by per video frame. Audeering's *openSMILE* (Schuller et al., 2009; Eyben et al., 2010) toolkit was used to extract 34 different audio features, including pitch, intensity, speech rate, and MFCCs per frame. Finally, Amazon Transcribe was used to transcribe speech and *deep-disfluency* Hough and Schlangen (April, 2017) was used to extract disfluent words in the form of transcribed text. Similar to Kaushik et al. (2021), disfluencies like edits, repair, reparandum, and interregnum word tokens were further identified.

Table 1: Text Encoding Feature Descriptions

Feature	Type	Description
$is\_question$	bool	token is part of question
is_pause	bool	token is $\geq 0.398s$ pause within utterance
$curr\_sentence\_length$	int	number of words in current sentence
$speech\_rate$	float	words/min. of current sentence
is_edit_word $is\_reparandum$ $is\_interregnum$ is_repair	bool	generated by Deep Disfluency

Using the output of Amazon Transcribe, each participant's text was divided into a sequential list of tokens, where a token could be a spoken word or a period of silence. For each token, we extracted 8 features, as shown in Table 1.

Since Amazon Transcribe's output was tagged with timestamps, we were able to align the text, audio, and video features. With missing data eliminated or smoothed out by inserting the averages of data in nearby frames, the audio and visual feature vectors for each word token were then taken to be

the averages of all the frames within the token's given time-span. Participant confusion labels over each time-span were finally collapsed to the most-occurring label for each word/token. Participant confusion labels were "smeared" (or duplicated) over frames according to their time-step, such that each frame was associated with the confusion label that the participant had selected at that time marker.

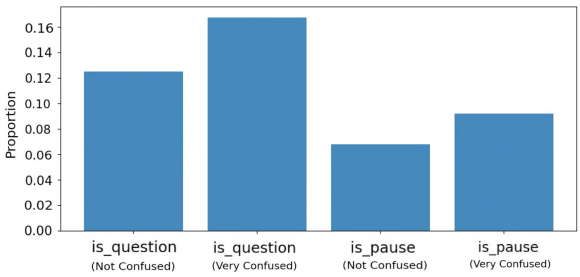

Figure 5: **Text Features:** The y-axis is the proportion of tokens that were: (a) a part of a spoken question, or (b) a distinct pause in the participant's speech. Observe that a higher proportion of tokens are part of a question or a pause when the participant is highly confused.

3.3 Exploration of Hidden Markov Models

We explored Hidden Markov Models (HMMs) because of their interpretability and applicability to sequential data. The HMM relies on the Markov assumption, which means that the state of the system at time step i is only dependent on the state of the system at time step $i-1$. Experimenting with different temporal increments based on video frames or word tokens, our model was unable to accurately predict confusion. This left us with a trade-off: either use all our frame-by-frame data and have the time increment be so small that the HMM yields a diagonal-heavy transition matrix or have longer increments but make our dataset prohibitively small by averaging over longer intervals.

3.4 The Neural Modeling Approach

We designed recurrent neural networks (RNNs) given their ability to extract temporal dependencies inherent to sequences (Schäfer and Zimmermann, 2006; Ororbia II et al., 2017). In essence, RNNs are stateful ANNs that "remember" information in prior time-steps $< t$ when processing data at t.

While taking a neural engineering approach offers a great deal of flexibility in terms of the type of architecture that one might design to process streams of different modalities (meaning there are many possible model designs we could craft), in this work, we take a simple approach. For each data modality, we crafted one RNN modality-processing model that specifically implements $p(y_t|\mathbf{x}_0^m, \mathbf{x}_1^m, ..., \mathbf{x}_t^m; \Theta_m) = f^m(\mathbf{x}_t^m; \Theta_m)$ where y_t is the (integer) confusion label[2] at time t and $\mathbf{x}_t \in \mathcal{R}^{O\times 1}$ is the specific feature vector (with O feature values) for modality m, where $m = \{\text{vis}, \text{aud}, \text{txt}\}$ (*vis* means visual, *aud* means audio, and *txt* means text/symbols) and Θ_m contains all of the learnable weight parameters. Concretely, any modality-processing RNN with H hidden neurons is specified by the dynamics:

$$\mathbf{h}_t = \phi_h(\mathbf{W}^m \cdot \mathbf{x}_t^m + \mathbf{V}^m \cdot \mathbf{h}_{t-1} + \mathbf{b}^m) \quad (1)$$

$$\hat{\mathbf{y}}_t = \phi_o(\mathbf{U}^m \cdot \mathbf{h}_t + \mathbf{c}^m) \quad (2)$$

where $\phi_(v) = \max(0, v)$ is the linear rectifier used for the hidden layer activation function, $\phi_(\mathbf{o}) = \exp(\mathbf{o}) / \sum_j \exp(\mathbf{o})[j]$ is the softmax used for the output layer, $\cdot$ denotes matrix-vector multiplication, and $\odot$ denotes the Hadamard product.

$\mathbf{W}^m \in \mathcal{R}^{H\times O}$ is the input-to-hidden weight matrix, $\mathbf{V}^m \in \mathcal{R}^{H\times H}$ is the recurrent weight matrix, and $\mathbf{U}^m \in \mathcal{R}^{O\times H}$ is the output/feature emission matrix while $\mathbf{b}^m \in \mathcal{R}^{H\times 1}$ and $\mathbf{c}^m \in \mathcal{R}^{O\times 1}$ are bias vectors. The RNN weight parameters $\Theta_m = \{\mathbf{W}^m, \mathbf{V}^m, \mathbf{U}^m, \mathbf{b}^m, \mathbf{c}^m\}$ are initialized using a scaled, centered Gaussian distribution and parameters are fit data using backpropagation through time to calculate the gradients of the cost function $\mathcal{L}(\hat{\mathbf{y}}_t, \mathbf{y}_t) = \sum_{t=1}^{T} -\sum_j (\mathbf{y}_t \odot \log(\hat{\mathbf{y}}_t))[j]$. The resulting $\frac{\partial \mathcal{L}(\hat{\mathbf{y}}_t, \mathbf{y}_t)}{\partial \Theta^m}$ (the partial derivatives) is used to adjust Θ using stochastic gradient descent based on the Adam update rule (Kingma and Ba, 2014).

Given the three modality-processing RNNs we trained, i.e., $f^{vis}(\mathbf{x}_t^{vis}, \Theta_{vis})$, $f^{aud}(\mathbf{x}_t^{aud}, \Theta_{aud})$, $f^{txt}(\mathbf{x}_t^{txt}, \Theta_{txt})$, final label predictions were made using a late-fusion aggregation scheme (Snoek et al., 2005). In other words, we computed the final predicted label y_t as follows: $y_t = \arg\max(\alpha_{vis}\mathbf{y}_t^{vis} + \alpha_{aud}\mathbf{y}_t^{aud} + \alpha_{txt}\mathbf{y}_t^{txt})$, which returns the index of class within the average of the three modal probability distributions. Importance weights α_{vis}, α_{aud}, and α_{txt} were set to 1.0, which means we assume equal weight per modality.

4 Results

4.1 Recurrent Neural Modeling Results

Our RNN modality-processing system was trained only on single modalities, with the final predicted

[2]We further encode this as a one-of-C binary vector $\mathbf{y}_t \in \mathcal{R}^{C\times 1}$, where C is the number of confusion levels/classes.

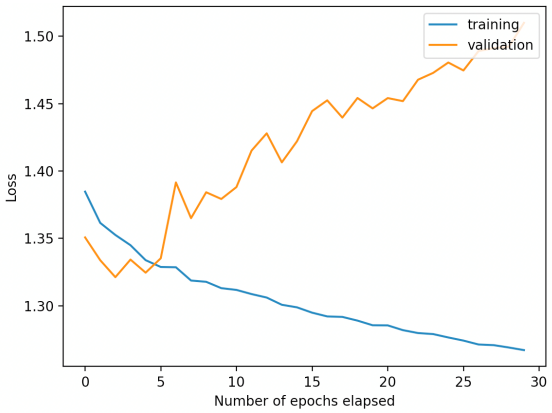

Figure 6: Training and validation loss (for the video modality) of the RNN system. It is evident that after 5 epochs, the model begins to severely overfit the training data, as the training loss continues to decrease while the validation loss begins to increase.

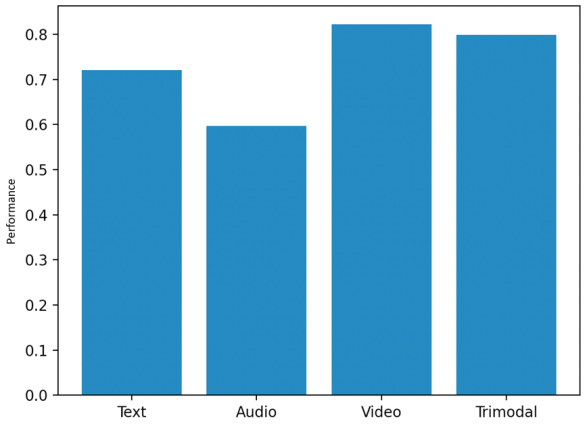

Figure 7: Unimodal and trimodal RNN model performance: the video-only model performs the best, followed by the trimodal late-fusion, text-only, and audio-only models.

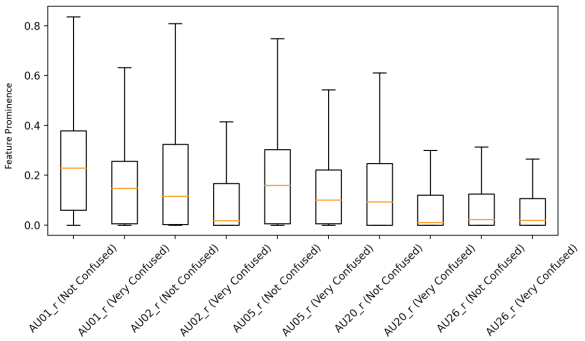

Figure 8: Box plots for select video features where the y-axis is the facial action unit reading produced by OpenFace: a 0-1 scale quantifies how heavily a facial action unit is being produced by a participant.

label y_t aggregated through the late-fusion scheme described above. We had 20 participants and held out one randomly selected female and male participant to validate model performance. In Figure 6, we present training and validation loss curves (total loss value plotted against epoch of training) for the model trained on video features only. This model performed better than the unimodal text and audio models, as well as the late-fusion trimodal model.

An RNN trained on only video features was able to achieve the lowest loss and best accuracy performance, suggesting video data conveyed the most meaningful knowledge about the confusion state. However, this model begins to overfit the training data around epoch 5 (Figure 6), at which point the training loss continues to decrease while the validation loss begins to increase. Changing parameters like the number of hidden neurons did not reduce the model overfitting though future work will investigate regularization schemes. Given our small dataset, the model appears to struggle to generalize to the two unseen participants. When we early-stop the training after 5 epochs to combat overfitting, we obtain the validation accuracy values for each uni-modal model shown in Figure 7.

4.2 Modality-Based Data Analysis

To inspect which features were possible predictors of confusion, we created box plots and bar charts to examine the distribution of feature values from participants while in the *Not Confused* state versus the *Very Confused State*. The features examined in this analysis were selected based on which had the highest difference in median value between the *Not Confused* and *Very Confused* states. In Figure 5, observe that some results make sense: participants are more likely to pause and ask questions when they are very confused versus when they are not confused at all. The facial action units are shown in Figure 8 and in Table 2. Intuitively, it makes sense that these facial action units are tied to confusion.

AU01	Inner brow raiser
AU02	Outer brow raiser
AU05	Upper lid raiser
AU20	Lip stretcher
AU26	Jaw drop

Table 2: The facial action units displayed in Figure 8.

Some analysis results, in contrast, are more surprising: we found nearly no difference between the distributions of the audio features extracted in the *Not Confused* versus the *Very Confused* states. This suggests that prosodic features are potentially less effective predictors of confusion in this study.

5 Discussion

Given the results of the previous section, we discuss the contributions of our work driven by our initially presented questions. Specifically, we make the following contributions which we next state as answers to our original research questions:

RQ 1: The annotation method used by Kaushik et al. (2021) involved participants marking their confusion level for every 30-second block. We improved upon this approach by implementing the *Confuse-o-Meter* website, which allowed participants to continuously annotate their confusion levels. This method for annotation was found to provide a richer dataset in which we were able to obtain confusion labels for every time-step of data.

RQ 2: The RNN results showed that we were able to build a model that could relatively accurately classify confusion in the test set participants.

RQ 3: There were inconclusive results on which facial, audio, and language features were the best predictors of confusion because different methods yielded conflicting results. However, based on our limited results, we reason that the following features may be linked with confusion: text disfluencies, pauses, questions, AU01 (inner brow raiser), AU02 (outer brow raiser), AU05 (upper lid raiser), AU17 (lid tighten), AU20 (lip stretcher), AU23 (lip tighten), and AU26 (jaw drop).

The main limitation of our work is the size of the collected dataset – with only 20 participants, it makes sense that our models, particularly the highly nonlinear RNN system, overfit to the training samples. For any choice of two participants, it is unlikely that a model trained on 18 other participants would generalize to the test participants since confusion is a complicated emotion and not all humans display it the same way. It would take a larger dataset in order to generalize to the broader population. Additionally, our models predicted the *Not Confused* states more often than the *Confused* states. The distribution of our confusion dataset is similarly unbalanced, as seen in Figure 9.

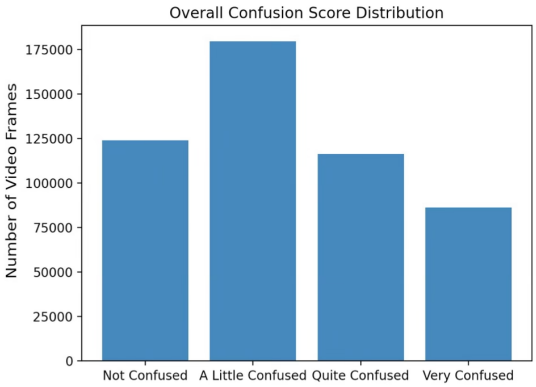

Figure 9: The confusion label distribution indicates that participants generally spent more time in the not confused states as opposed to the confused states.

6 Conclusions

In this study, our goal was to build a model that was capable of accurately predicting confusion and to understand which text, audio, and video features were accurate predictors of confusion. Given that the RNN has low interpretability, we utilized statistical methods to accomplish the latter half of this goal. Furthermore, we improved upon previous methods of data collection to allow for continuous annotation of confusion states. This design choice provided us with a more precise multimodal dataset with rich confusion reference labels across time. To computationally model the predictive label distributions and perform confusion classification, we constructed a computational model based on recurrent neural networks (RNNs), which lack interpretability but proved to be reasonably accurate even with our limited data. Future work will include generalizing our RNN computational model further to better handle the different modalities (in an intermediate modality fusion scheme as in Ororbia et al. (2019)) found within our dataset, as opposed to our current method of taking the (late-fusion) weighted consensus of three separately trained modality-processing RNNs. In addition, another future next step would be to repeat our study to collect a larger dataset that better represents the general population. This may also reduce the overfitting observed in our predictive confusion models. Additional research could investigate dimensionality reduction techniques and alternative forms of statistical analysis to explore measured features in our data.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Award No. IIS-1851591 and DGE-2125362. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

American Psychology Association. 2021. Confusion. https://dictionary.apa.org/confusion.

Thushari Atapattu, Katrina Falkner, Menasha Thilakaratne, Lavendini Sivaneasharajah, and Rangana Jayashanka. 2020. What do linguistic expressions tell us about learners' confusion? a domain-independent analysis in moocs. *IEEE Transactions on Learning Technologies*, 13(4):878–888.

Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Facial Expression Recognition and Analysis Challenge, IEEE International Conference on Automatic Face and Gesture Recognition*.

Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, , and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face and Gesture Recognition*.

Paul Ekman and Wallace V. Friesen. 1976. Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, 1(1):56–75.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. opensmile – the munich versatile and fast open-source audio feature extractor. In *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*, pages 1459–1462.

Julian Hough and David Schlangen. April, 2017. Joint, incremental disfluency detection and utterance segmentation from speech. In *EACL 2017*, Valencia, Spain.

Nikhil Kaushik, Reynold J. Bailey, Alexander G. Ororbia, and Cecilia O. Alm. 2021. Elicitation of confusion in online conversational tasks. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association*, Brno, Czech Republic.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Alexander Ororbia, Ankur Mali, Matthew Kelly, and David Reitter. 2019. Like a baby: Visually situated neural language acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5127–5136.

Alexander G Ororbia II, Tomas Mikolov, and David Reitter. 2017. Learning simpler language models with the differential state framework. *Neural computation*, 29(12):3327–3352.

Anton Maximilian Schäfer and Hans Georg Zimmermann. 2006. Recurrent neural networks are universal approximators. In *International Conference on Artificial Neural Networks*, pages 632–640. Springer.

Björn Schuller, S. Steidl, and Anton Batliner. 2009. The interspeech 2009 emotion challenge. In *Interspeech 2009, 10th Annual Conference of the International Speech Communication Association*, Brighton, UK.

Zheng Shi, Ya Zhang, Cunling Bian, and Weigang Lu. 2019. Automatic academic confusion recognition in online learning based on facial expressions. In *2019 14th International Conference on Computer Science Education (ICCSE)*, pages 528–532.

Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. 2005. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402.